2.1: Visualizing Variation in Numerical Data

Definition.
The **distribution of a sample** of data is a way of organizing the data by recording the

- values that were observed, and
- the frequencies of these values.

Example. Below are the number of goals scored by first year NCAA female soccer players in Division III in the 2016-17 season:

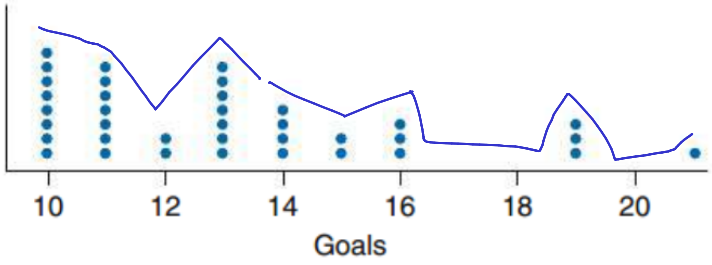
$$11, 14, 16, 13, 13, 10, 13, 11, 16, 21, 13, 19, 10, 10, 14, 13, 10, 13,$$

$$15, 10, 15, 13, 11, 19, 11, 11, 16, 10, 12, 11, 14, 11, 10, 14, 10, 19, 12$$

The **distribution** lists the values *and* the frequencies:

Value	Frequency
10	8
11	7
12	2
13	7
14	4
15	2
16	3
17	0
18	0
19	3
20	0
21	1

A **dotplot** represents the data by using a dot where each value occurs:

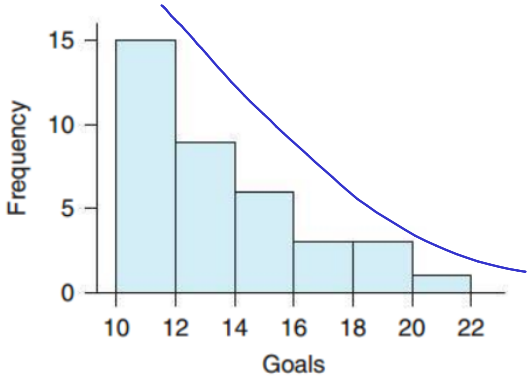


▲ **FIGURE 2.2** Dotplot of the number of goals scored by first-year women soccer players in NCAA Division III, 2016–17. Each dot represents a soccer player. Note that the horizontal axis begins at 10.

Histograms:

A **histogram** represents the data by using bars to indicate how much data lies in each *bin* (also called *interval* or *class*):

▶ **FIGURE 2.3** Histogram of number of goals for female first-year soccer players in NCAA Division III, 2016–17. The first bar, for example, tells us that 15 players scored between 10 and 12 goals during the season.

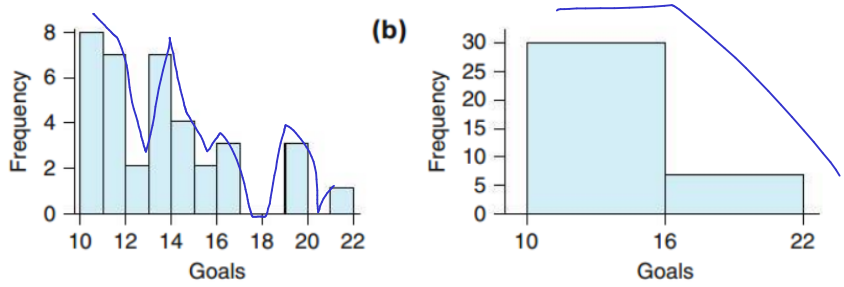


Q: Where do we place data points that lie on a boundary?

Note: Bin size plays a significant role in how the data is represented in a histogram. A bin width that is:

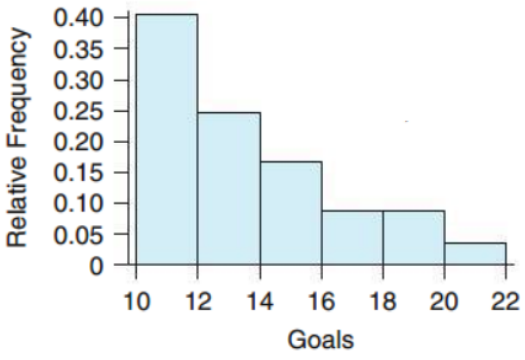
- too narrow shows too much detail.
- too wide hides detail.

▶ **FIGURE 2.4** Two more histograms of goals scored in one season, the same data as in Figure 2.3. **(a)** This histogram has narrow bins and is spiky. **(b)** This histogram has wide bins and offers less detail.



A **relative frequency histogram** changes the units on the vertical axis to represent relative frequencies:

▶ **FIGURE 2.5** Relative frequency histogram of goals scored by first-year women soccer players in NCAA Division III, 2016–17.



Stemplots:

Definition.
A **stemplot** divides each observation into a *stem* and *leaf*. The **leaf** is the last digit in the observation, and the **stem** contains all the digits preceding the leaf.

Example. A collection of college students who said that they drink alcohol were asked how many alcoholic drinks they had consumed in the last seven days. Their answers were:

$$1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 8, 10, 10, 15, 17, 20, 25, 30, 30, 40$$

Stem	Leaves
0	111112223333345556668
1	0057
2	05
3	00
4	0

Example. Below is a stemplot for exam grades. How many grades are between 40% and 59%?

Stem	Leaves
3	8
4	
5	
6	0257
7	00145559
8	0023
9	0025568
10	00

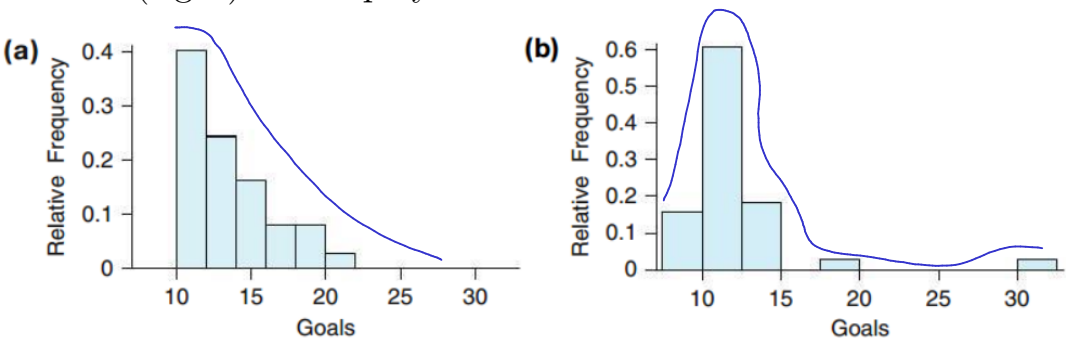
2.2: Summarizing Important Features of a Numerical Distribution

Definition.
When examining a distribution:

- the **center** represents the typical or most common values, and
- the **spread** represents the variability in the data.

Example. Below are the histograms containing the number of goals scored by first year NCAA female (left) and male (right) soccer players in Division III in the 2016-17 season:

▶ **FIGURE 2.9** Distributions of the goals scored for **(a)** first-year women and **(b)** first-year men in Division III soccer in 2017.



– Are their any notable differences in the shapes?

Yes; female goals are highest in 10-12 range, then taper off
male goals are also highest in 10-12, but also are on either side

– What is the approximate center for each distribution?

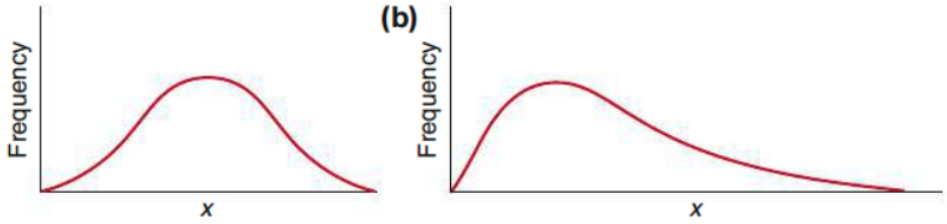
Female: approximately 15 goals
Male: approximately 12 goals

– How do the spreads compare?

Distribution of male goals is more spread out

Three basic characteristics to consider when examining a distribution's shape:

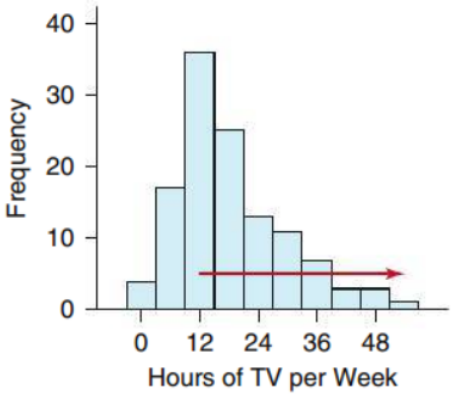
1. Is the distribution symmetric or skewed?
2. How many "mounds" appear?
3. Are unusually large or small values present?



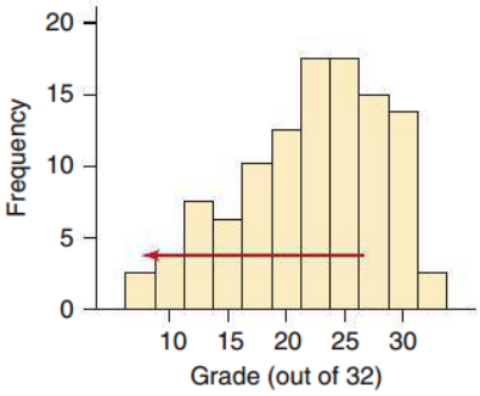
▶ **FIGURE 2.10** Sketches of **(a)** a symmetric distribution and **(b)** a right-skewed distribution.

Definition.

- A **right-skewed distribution** has a "tail" that extends towards the right.
- A **left-skewed distribution** has a "tail" that extends towards the left.
- A **symmetric** distribution has "tails" of approximately equal size.



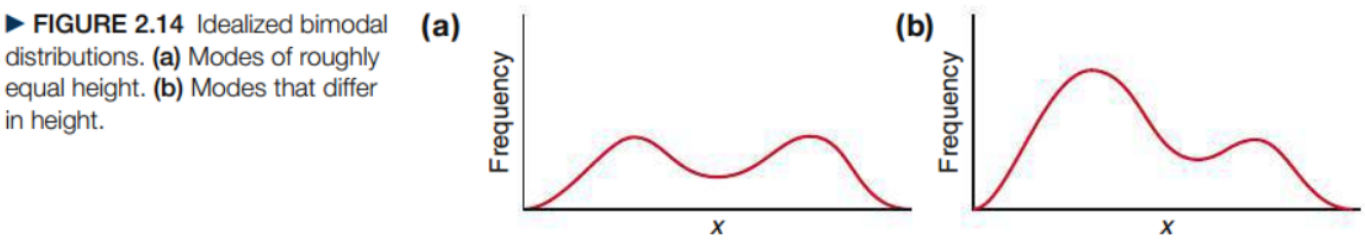
▲ **FIGURE 2.12** This data set on TV hours viewed per week is skewed to the right. (Source: Minitab Program)



▲ **FIGURE 2.13** This data set on test scores is skewed to the left.

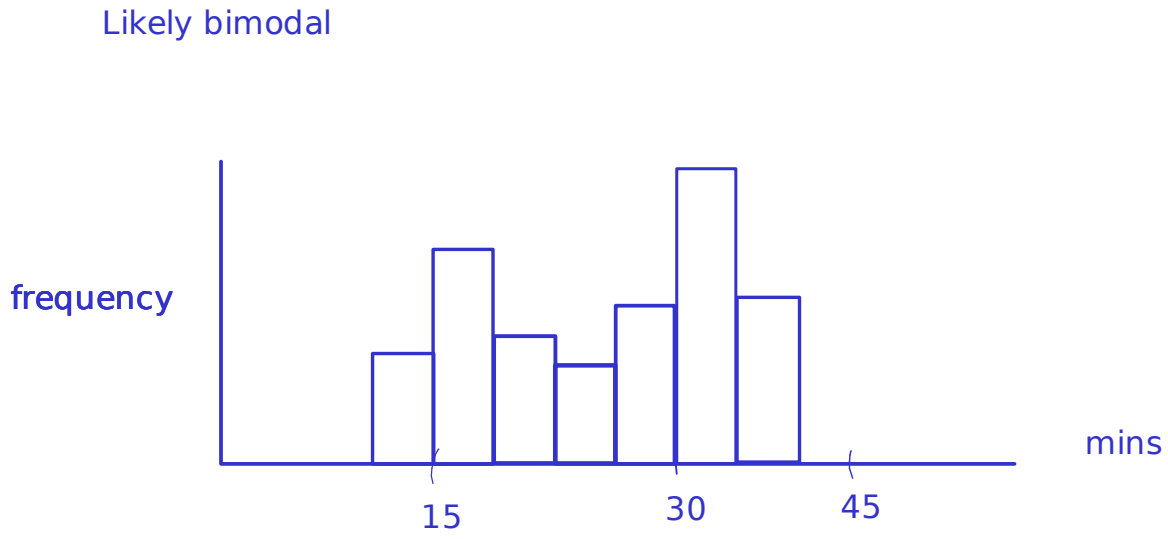
Definition.

- A **unimodal distribution** has data grouped in a single "mound",
- a **bimodal distribution** has data grouped in two "mounds", and
- a **multimodal distribution** has data grouped in more than two "mounds".



▶ **FIGURE 2.14** Idealized bimodal distributions. **(a)** Modes of roughly equal height. **(b)** Modes that differ in height.

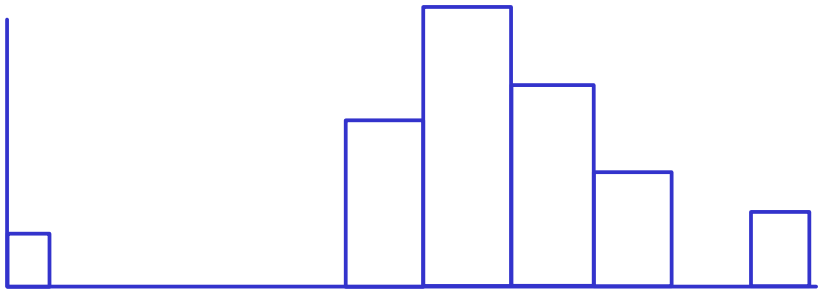
Example. In a 5k/10k race where all the runners start at the same time, what do we expect the shape of the distribution of the finishing times will look like?



Definition.
An **outlier** is an extreme value in a distribution of data. Outliers don't fit the pattern of the rest of the data.

Example. Consider the distribution of exam grades. What are possible explanations of any outliers?

-Someone didn't take the exam
-Perhaps someone took the class before



Definition.
The most frequently occuring value is called the **mode**.

Why might the mode not be a reliable measure of center for numerical data?

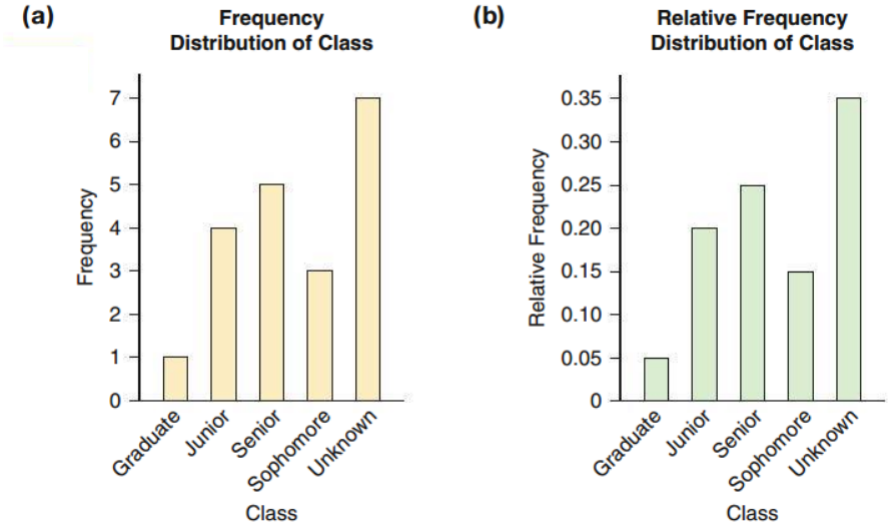
Not representative of ALL the data
Might move around if new data points are added

2.3: Visualizing Variation in Categorical Variables

Definition.
A **bar chart** (also bar graph or bar plot) shows a bar for each observed category where the height of the bar is proportional to the frequency of that category.

Example. A summer introductory statistics course at UCLA has the following distribution of students across different years:

Class	Frequency
Unknown	7
Freshman	0
Sophomore	3
Junior	4
Senior	5
Graduate	1
Total	20



Bar Charts vs. Histograms:

- Bar charts are for categorical data
- Histograms are for numerical data

	Histogram	Bar Chart
Bars:	Should touch	May or may not touch
Bar width:	Corresponds to bin width	Can be any width (consistent)
Horizontal labels:	Numerical order	No inherent order

- A **Pareto chart** is a bar graph with bars arranged from tallest to shortest.

Definition.
A **pie chart** is a circle divided up into pieces where each area is proportional to the relative frequency of the category it represents.

Example.

Class	Frequency
Unknown	7
Freshman	0
Sophomore	3
Junior	4
Senior	5
Graduate	1
Total	20